IJCTA, 9(34) 2016, pp. 475-488
© International Science Press

Identification of Rare Genetic Disorder from Single Nucleotide Variants Using Supervised Learning Technique

Sathyavikasini K[1*] and Vijaya M S[2]

ABSTRACT

Muscular dystrophy is a rare genetic disorder that affects the muscular system which deteriorates the skeletal muscles and hinders locomotion. In the finding of genetic disorders such as Muscular dystrophy, the disease is identified based on mutations in the gene sequence. A new model is proposed for classifying the disease accurately using gene sequences, mutated by adopting positional cloning on the reference cDNA sequence. The features of mutated gene sequences for missense, nonsense and silent mutations aims in distinguishing the type of disease and the classifiers are trained with commonly used supervised pattern learning techniques.10-fold cross validation results show that the decision tree algorithm was found to attain the best accuracy of 100%. In summary, this study provides an automatic model to classify the muscular dystrophy disease and shed a new light on predicting the genetic disorder from gene based features through pattern recognition model.

Keywords: cDNA, Codon, Codon Usage Bias, Positional Cloning, RSCU

1. INTRODUCTION

The majority of hereditary disorders place a significant burden on the families immortalizing the condition for the lack of effective treatment [1]. Muscular dystrophies are such trait caused by mutations in the gene sequences. Muscular dystrophy (MD) is a cluster of successive muscle disorders stimulated by mutations in genes that encode for proteins that are vital for regular muscle function [2, 3]. The results of muscle biopsy, electromyography, electrocardiography and DNA analysis aids in diagnosing muscular dystrophy. The disease should be diagnosed early and effectively to understand and improve the life of patients. Some forms of MD are Duchenne, Becker, Emery-Dreifuss, Limb-girdle, Facioscapulohumeral, Myotonic and Charcot Marie Tooth disease [4].

Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMD) are caused by the mutations in the dystrophin gene. Dystrophin is the hefty human gene that is 2.5mb long and encompasses of 79 exons. When the effect of mutations is less in the dystrophin gene it results in Becker's muscular dystrophy [5, 6]. Emery-Dreifuss muscular dystrophy (EMD) can be affected in patients, typically in their childhood and in the early adolescent years with muscle contractures. The genetic changes in the Emerin (EMD) and Lamin A/C (LMNA) genes cause Emery- Dreifuss muscular dystrophy [7]. Limb-girdle muscular dystrophy (LGMD) can be seen in both boys and girls. Nearly 18 genes involved in the mutation of LGMD. The defects in LGMD show a related distribution of muscle weakness that has an effect on both upper arms and legs. Charcot Marie tooth disease (CMT) includes a number of disorders with an assortment of symptoms.

The Single Nucleotide Variants (SNV) causes a distinct variation in the genetic code of the DNA sequence. These changes are termed as mutations. Mutations in the gene sequence make a permanent change in the

[1] PhD Research Scholar, PSGR Krishnammal College for Women, Coimbatore 641004, India, *E-mail: Mail2sathyavikashini@gmail.com*

[2] Associate Professor, PSGR Krishnammal College for Women, Coimbatore 641004, India, *E-mail: msvijaya@psgrkc.com*

DNA sequence that clearly roots to genetic disorder. The impact of the SNV on the gene sequence modifies the function of the gene. Substitution is an exchange of one base to another, such as swapping a base from A to G. SNV's may be synonymous or non-synonymous. Missense and non sense are the non synonymous single nucleotide variants where a single change in the gene alters the amino acid in the sequence [8, 9]. Missense mutations are the substitution in a codon that encodes a different amino acid and alters the protein [10]. Nonsense mutations are those where the protein attains to stop codon when a change occurs in the DNA sequence.

Synonymous mutations are the silent mutations that the variant will not show amend in the amino acids. Silent mutations are a change in codon that encodes for the same amino acid and therefore the translated protein is not modified [11]. In detecting the type of disease it is necessary to consider the silent mutation as the changes can affect protein folding and function. Even though several codons encode for the same amino acid their frequency will vary and this is referred as codon bias. The increase in the number of the same nucleotides in a location is termed as duplications. Deletions are the mutations when a base or an exon is deleted from a sequence the mutations. [12].

Muscle biopsy and DNA testing are in progress for diagnosing muscular dystrophy [13]. An initial step in examining muscular dystrophy is through genetic testing. The advantage of performing genetic testing over muscle biopsy is that in genetic testing, the blood sample is enough to spot the alteration in the genes whereas the part of the tissue is required to perform the muscle biopsy. [14]. Gene therapy helps in knowing the exact mutation in the DMD gene and direct sequencing aids in identifying missense, nonsense, insertions, deletions and splicing mutations [15, 16].

All sort of mutations cannot be identified out using Multiplex Ligation-dependent Probe Amplification (MLPA) and in some cases, the results would be negative [17]. In the case of DMD, SNV's are detected by means of Sanger's full gene sequencing, which is performed by direct sequencing methodology. The direct sequencing analysis is considered to be laborious, expensive and time-consuming [18, 19]. PCR is now a common and often indispensable technique used in medical and biological research labs in the diagnosis of hereditary diseases [20, 21].

The laboratory methods are facing challenges in analyzing the gene sequences to detect the genetic disorder. Therefore, the process should be automated through the computational methods and disease should be identified efficiently.

Classification of Facioscapulohumeral muscular dystrophy (FSHD) disease is done by monitoring of expression levels. Usually, microarray gene expression analysis is mainly focused to cancer diseases. In the paper [22], the authors proposed an approach to classifying the types of Facioscapulohumeral muscular dystrophy (FSHD). A model is created using Support vector machine to classify the types of FSHD.

The authors Catherine T. Falk, James M. Gilchrist [23] developed a model using neural networks to identify whether the patient is affected from Limb Griddle muscular dystrophy (LGMD). The data based on the patients' family details are collected. The classification of disease status is made using the neural network and achieved an accuracy of 98%.

The authors in [24] constructed a protein – protein interaction network to classify the sub types of muscular dystrophy through machine learning techniques. Microarray gene expression datasets are analyzed and the protein data and their interaction data are collected and a network is constructed to classify the sub types. Multi class support vector machine is applied for the classification of six sub- types of muscular dystrophy.

Some of the limitations of micro array data to classify all forms of muscular dystrophy are the cDNA probes plotted on the microarrays do not cover all of the genes expressed in skeletal muscle, the properties of probe cDNAs have not been not well-characterized, homologous genes of each target gene may crosshybridize with the probes and because relatively large amounts of RNA are required, each microarray analysis has required pooled RNA samples from several patients [25, 26].

The authors in [27] proposed a model to classify the types of Human Leukocyte Antigen (HLA) gene into different functional groups by choosing the codon usage bias as input. In their work, they converted the gene sequence into 59 vector elements by calculating the RSCU values for the gene sequence. A model was created using Support vector machine and achieved an accuracy rate of 99.3 percent.

The authors C.M.Nisha, Bhasker Pant, and K. R. Pardasani proposed an approach based on codon usage pattern to classify the type of Hepatitis C virus (HCV) that are the primary reason for the liver infection. To classify the subclass of its genotype a model was created using codon usage bias as input to multi class SVM [28].

The classification of muscular dystrophy continues to evolve with the advances in understanding of their molecular genetics. Huge number of muscular dystrophy related faulty genes and proteins are identified, but no successful treatments are known for many of its sub-types. The proportion of mutations in deletions, duplications and point mutations differs in each type of disease and the present methods cannot handle the entire mutational spectrum in a single platform. However, it is essential to look into the accurate mutation site and to predict the disease.

In the above mentioned literatures the disease classification is done for only some kind of muscular dystrophy diseases. The classification was performed with the data such as micro array gene expression data, protein interaction data and with family details. The synonymous variants were captured with the RSCU values that helped in identifying the virus or in classification of gene sequences. Hence, it is motivated that the classification of disease can also be carried out by modeling both synonymous and non synonymous mutations using diseased gene sequences.

As Muscular dystrophy is a genetic disorder, it is imperative to identify from the mutations in the gene sequences. Synonymous and non – synonymous SNV's must be considered to predict the disease efficiently. Hence, in this paper the disease is predicted from the mutated gene sequences by building a model using supervised learning algorithms for all types of single nucleotide variants.

In this research work diverse features are designed to propose a new model and an integrated approach is demonstrated based on computational intelligence technique to detect major five forms muscular dystrophy with cloned gene sequences as input. Features about missense, non-sense mutations and silent mutations in gene sequences are identified and a model is generated using supervised learning technique.

2. MATERIALS AND METHODS

The gene sequences and its pattern vary in every human. Also the pattern gets altered when mutations occur in the chromosome. The principal focus of this research is to identify discriminative features and to provide an efficient machine learning solution for predicting the type of muscular dystrophy disease with the silent mutations. Multi-class classification is formulated through data modeling of gene sequences. The synthetic mutational gene sequences are generated as the diseased gene sequences are not readily available for this complicated disease. Five types of muscular dystrophy namely DMD, BMD, EMD, LGMD and CMT have been considered for building the disease prediction model.

Disease Identification Model

The development of Muscular dystrophy disease Identification model comprises of phases such as mutational gene sequence generation, feature identification and extraction, building the model and classification. The framework of the proposed model is illustrated in Fig. 1

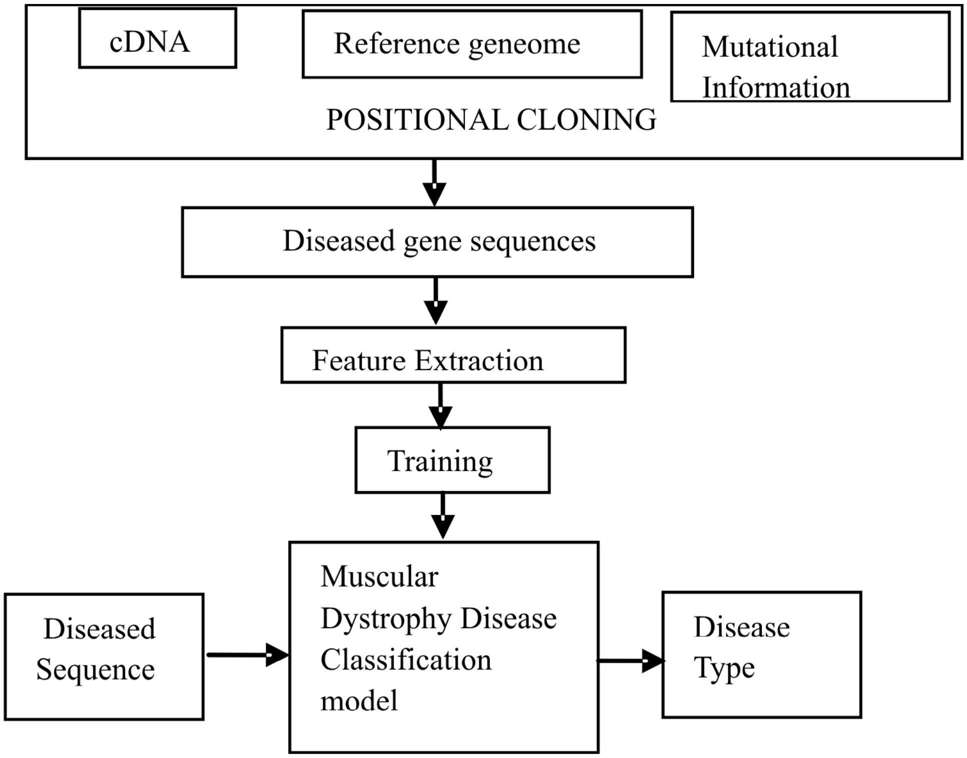

Figure 1: Workflow of Muscular dystrophy Disease Identification model

Positional cloning

Positional cloning is a traditional approach to recognize the disease based on its location on the chromosome. Positional cloning aids in disease identification even when minute information is known about the molecular basis of the trait. The first gene cloned by positional cloning methodology was the dystrophin gene to diagnose DMD and hence the same approach is applied in this work to generate mutated gene sequences, encoding all the required genes.

Mutated gene sequences are generated by this approach based on the mutation and its location on the chromosome. The information on the position of mutations in the gene sequences is available in HGMD[1] (Human Gene Mutation Database) [28] is a collection of data on germ-line mutations in genes with their human hereditary disease which are grasped from various literatures. The open version of HGMD[1] is available for non –commercial purpose for the registered users in Educational institutions/non-profit organizations.

The positional change of the nucleotide is done in cDNA sequence against the reference gene sequence and the new mutated gene sequences for muscular dystrophy are generated through R script. The cDNA sequence and the reference sequence are first stored as text files. Using the Stringreplace() function from the stringi library the required position is to be altered is identified and replaced with the nucleotide specified in the nucleotide change column of HGMD database.Five types of mutations have been considered for generated for generating mutated sequences. Using the traditional positional cloning approach the mutated sequences are generated and stored as fasta files.Consider the missense mutational information for the EMD phenotype from the Emerin gene such as nucleotide change is 2 T>C which indicates in the position 2 the nucleotide changes from T to C alters the protein from Met to thr.

Disease gene datasets

The genes associated with diseases are examined. There are about fifty five genes evacuated with five type of muscular dystrophy. Table I summarizes genes associated with the disease. The mutational information is retrieved from the HGMD database using the gene information for the required phenotype. The corpus of data holds all types of mutated sequences such as Missense, Non sense, synonymous, Insertion and deletion mutations. A set of 30 mutated gene sequences for each disease is generated for all type of phenotypes. The dataset comprises of 150 mutated gene sequences combining all forms of muscular dystrophy is developed.

Table I
List of genes associated with different type of muscular dystrophy

Muscular dystrophy Disease	*Genes associated with the disease*
Duchenne muscular dystrophy	Dystrophin
Becker's muscular dystrophy	Dystrophin
Emery-dreifuss	Emerin ,LMNA/C
Limb griddle muscular dystrophy	ANO5, CAPN3, CAV3, DYSF, FKRP, FKTN, LMNA, MYOT, POMGNT1, POMT1, POMT2, SGCA, SGCB, S, GCD, SGCG, TCAP, TRIM32, TTN
Charcot marie tooth disease	AARS, AIFM1, BSCL2, DHTKD1, DNM2, DYNC1H1, EGR2, FGD4, FIG4, GARS, GDAP1, GJB1, HSPB1, HSPB8, INF2, KARS, KIF1B, LITAF, LMNA, LRSAM1, MED25, MFN2, MPZ, MTMR2, NDRG1, NEFL, PMP22, PRPS1, PRX, RAB7A, SBF2, SH3TC2, TRPV4, YARS

Feature Extraction and Training dataset

Change in the structure of sequence implies the cause of the disease. These structural changes can be captured as features for mutational sequence to learn the prediction model. The codon usage patterns are considered as the contributing features for representing silent mutations in the mutated gene sequences. Since codon usage patterns are diverse in different gene families, this feature input is a well-chosen descriptors for specifying different gene families for all types of diseases. Eighty eight evocative features for both the synonymous and non synonymous mutations are extracted and feature vectors are created for learning disease prediction model.

Features of Missense and Nonsense mutations

The missense and nonsense mutational features are based on annotation, structure and alignment of the diseased gene sequences. They are GeneID, Gene symbol, Chromosome number, Alteration type, Protein changed, Reference allele, Observed allele, Mutation position, Length of the sequence, Mutation start position, Mutation end position, Position of mutation in gene sequence, amino acid change leads to stop codon or not, stop codon, Position of start codon in cDNA sequence, position of stop codon in DNA sequence, the nucleotide composition of A, G, C, T, AT and GC component composition, Edit distance scores, PhredQuality scores, Substitution scores and RSCU values from 59 codons. The features are extracted from a set of 150 mutated sequences

These features are extracted from mutated gene sequences through R script. There are numerous packages available in R for bioinformatics applications that are downloaded from www.CRAN.org.

The attributes of gene sequences like Gene ID, Gene symbol Chromosome number are identified by using the biomart package in R. These annotation features are extracted using getgenes(id). The Gene ID is the NCBI gene identifier for the affected phenotype. Some examples are GeneID 1746 is for Dystrophin gene, 2010 for Emerin gene, 4000 for LMNA gene etc. The symbol of the gene associated to the cDNA sequence of the disease such as DMD, LMNA and SMCHT32 etc.

The alteration type such as missense, nonsense, silent, deletion and duplications are encoded to numeric values from 1 to 5. The reference allele is the actual protein that is present in the cDNA sequence file and the observed allele is the protein observed after alteration. To identify the reference allele and observed allele, the position of codon is to be identified from the mutated sequence file. The first step in finding the observed allele is to read the fasta file and split it into codons. The required codon is acquired and altered based on the position information of codon change. Seqinr, and Biostrings library are inquired for this work.

The Length of the sequence is captured using the Length() function The fasta file is converted to dataframe and the length of the sequence is determined. The position of mutation in the gene sequence is identified by blasting the mutated sequence against the reference gene sequence. Nucleotide blast is used to capture the position of gene sequence.

The base composition A, C, G,T are calculated to count the number of occurrences of the four different nucleotides (“A”, “C”, “G”, and “T”) in the sequence. The most fundamental properties of a genome sequence is its AT and GC content, GC content is the fraction of the sequence that consists of Gs and Cs, ie. The GC content can be calculated as the proportion of the bases in the genome that are Gs or Cs. That is,

$$\text{AT content} = (\text{number of As} + \text{number of Ts})*100/\ (\text{genome length})$$

$$\text{GC content} = (\text{number of Gs} + \text{number of Cs})*100/\ (\text{genome length})$$

The position of the Stop codon reveals the end of the coding part in the sequence. To find the position of start codon matchpattern() function is used. Alignment scores are considered as the important feature for disease prediction. The global pair wise alignment based on edit distance is done with the mutated sequence against with the reference cDNA sequence and the alignment scores are calculated using edit distance scoring method. The PhredQuality measures are calculated with the patternQuality and subjectQuality to examine the quality-based match and mismatch bit scores for DNA/RNA. The substitution scores are calculated by setting the error probability to 0.1. Table II depicts the missense and nonsense features from mutated gene sequences.

Table II
Mutational features of missense and nonsense mutations

Features	*Description*
Gene ID	Identifier of the gene taken from NCBI
Gene Symbol	Name of the gene involved
Chromosome Number	The chromosome involved in mutation
Alteration type	Mutation type such as missense, non sense, silent, deletion and duplication
Protein changed	Whether protein altered through mutation
Observed allele	The amino acid present in normal gene
Reference allele	The observed amino acid after mutation
Mutation Position	Position of alteration in cDNA sequence
Length	Length of the mutated gene sequence
Mutation start position	The starting position of alteration in cDNA sequence
Mutation end position	The position where the mutation ends in cDNA sequence
Position	Mutation Position in gene sequence is identified through nucleotide blast against reference gene sequence
Nucleotide Composition	Composition of A, C, G, T, AT, GC in mutated sequence.
Position of stop codon	Last position of stop codon ATG
Edit distance scores	Alignment scores using edit distance method
PhredQuality measures	Calculated with patternQuality and subjectQuality
Substitution scores	Calculated with the error probability set to 0 or 1
ConsensusStart	The starting position of conserved region
ConsensusEnd	The end position of the conserved region

Features of Silent Mutations

A codon is the triplet of nucleotides that code for a specific amino acid. Many to one relationship occur between the codon and amino acid. Many amino acids are coded by more than one codon because of the degeneracy of the genetic codes.A total number of codons in a DNA sequence counts to 64. Since methionine (ATG) and tryptophan (TGG) have only one corresponding codon, they are not counted and are eliminated from the analysis as their RSCU values are always equal to 1. The three stop codons (TGA, TAA, TAG) are also not included. Accordingly, the number of codons considered is 59. Therefore, irrespective of the size, the DNA sequence is converted to a feature vector of 59 elements.

The differences in the frequency of occurrence of synonymous codons are referred as codon usage bias. The formula for calculating RSCU can be explained as, the number of times a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias [26].The RSCU carries the value 1.00 if the codon usage bias of that particular codon is absent. If the codon is used less frequently than expected, the RSCU values tend to have the negative values. Following formula is used to calculate RSCU.

$$\text{RSCU} = X_{ij} / (1/n_i * S \{X_{ij}; j=1, n_i \})$$

where X_{ij} is the number of occurrences of the j^{th} codon for the i^{th} amino acid, and n_i is the number of alternative codons for the i^{th} amino acid. If the synonymous codons of an amino acid are used with equal frequencies, then their RSCU values are 1. The RSCU values are derived for 59 codons from each mutated gene sequence which forms a feature vector for classification task. Table III holds the sample RSCU values of 59 codons for a mutated gene sequence.

Table III
RSCU Values for 59 codons for a sample sequence

Codon	*Value*	*Codon*	*Value*	*Codon*	*Value*
AAA	1.05	CCC	0.97	GGC	0.92
AAC	0.812	CCG	0.12	GGG	0.75
AAG	0.948	CCT	1.64	GGT	0.64
AAT	1.18	CGA	0.87	GTA	0.81
ACA	1.52	CGC	0.54	GTC	0.93
ACC	0.76	CGG	0.66	GTG	1.40
ACG	0.24	CGT	0.63	GTT	0.85
ACT	1.48	CTA	0.73	TAC	0.61
AGA	1.84	CTC	0.87	TAT	1.38
AGC	0.99	CTG	1.41	TCA	1.23
AGG	1.42	CTT	1.03	TCC	0.91
AGT	1.36	GAA	1.22	TCG	0.14
ATA	0.52	GAC	0.81	TCT	1.33
ATC	1.10	GAG	0.77	TGC	1.16
ATT	1.36	GAT	1.18	TGT	0.833
CAA	0.87	GCA	1.23	TTA	0.71
CAC	0.86	GCC	1.18	TTC	0.64
CAG	1.13	GCG	0.15	TTG	1.23
CAT	1.14	GCT	1.42	TTT	1.63
CCA	1.25	GGA	1.67		

Building the model

The corpus holds 150 sequences of 5 types of Muscular dystrophy diseases such as Duchenne Muscular Dystrophy, Becker's Muscular Dystrophy, Emery Drefius Muscular Dystrophy, Limb Griddle Muscular Dystrophy and Charcot Marie Tooth Disease. A training set with 150 feature vectors has been created and for each feature vector the class label is assigned from 1 to 5 indicating the five types of muscular dystrophy diseases. The features obtained from each mutated gene sequence forms a feature vector for classification task.

The standard supervised learning techniques, namely Naïve Bayes Classifier, Decision tree induction and artificial neural network have been used to learn and build the classifiers. Independent trained models have been used for predicting the type of muscular dystrophy disease for single nucleotide variants. The performance of trained models is evaluated using 10-fold cross validation and measured in terms of classification accuracy. The prediction accuracy is calculated with the number of correctly classified instances in the test dataset against the total number of test cases.

3. RESULTS

Five types of muscular dystrophy disease categories - Duchenne muscular dystrophy, Becker's muscular dystrophy, Emery-Dreifuss, Limb-girdle muscular dystrophy and Charcot marie tooth disease are taken into account to implement a multi class classification model. Mutated gene sequences are generated and the features of missense, nonsense and silent mutations are extracted from the corpus of data. Annotation, structure and alignment features count to twenty nine for missense and nonsense mutations. As silent mutations plays a major role in detecting the synonymous variants of genetic disorder it is important to detect silent mutations from gene sequences. The Relative synonymous codon usage (RSCU) values are calculated from fifty nine codons are taken as features for silent mutations and feature vectors are designed. Standard supervised learning techniques including decision tree, artificial neural network, naïve bayes are utilized and a muscular dystrophy disease classification model is developed using R, which is an open source software environment for statistical computing. The average accuracy of the classifiers is evaluated using 10-fold cross validation and the performance of the model is evaluated. From the results, it is observed that decision tree algorithm attains a high accuracy value of 100% for the developed training data set. The results of the experiments are summarized in Table IV and Table V. Table VI gives the comparative analysis of the existing and proposed work.

Table IV
Predictive performance of the classifiers

Evaluation criteria	*Classifiers*		
	NB	*ANN*	*Decision Tree*
Kappa Statistic	0.7804	0.9117	1
Correctly classified instances	120	138	150
Incorrectly classified instances	30	12	150
Prediction accuracy	80%	92%	100%

The existing approaches either classify the gene or the disease with the micro array, protein or family details data. The classification was done only for either synonymous or non synonymous type of SNV. The proposed approach can classify five types of muscular dystrophy from mutated gene sequence as input for both SNV's with 100% accuracy.

Table V
Statistics of classifier by its class

Class	*Classifier*	*Sensitivity*	*Specificity*	*Pos Pred Value*	*Neg Pred Value*	*Prevalence*	*Detection Rate*	*Detection Prevalence*	*Balanced Accuracy*
1	**ANN**	0.37	0.80	0.37	0.80	0.24	0.09	0.24	0.59
	NB	0	0.97	0	0.98	0.012	0	0.025	0.49
	DT	1	1	1	1	0.22	0.22	0.22	1
2	**ANN**	0.2	0.85	0.2	0.85	0.16	0.032	0.16	0.52
	NB	0.18	0.78	0.18	0.79	0.20	0.04	0.21	0.48
	DT	1	1	1	1	0.22	0.22	0.22	1
3	**ANN**	0.3	0.78	0.3	0.78	0.24	0.071	0.24	0.54
	NB	0.24	0.63	0.23	0.64	0.31	0.07	0.33	0.43
	DT	1	1	1	1	0.22	0.22	0.22	1
4	**ANN**	0.2	0.80	0.2	0.80	0.19	0.039	0.19	0.50
	NB	0.25	0.76	0.26	0.75	0.25	0.06	0.24	0.51
	DT	1	1	1	1	0.18	0.18	0.18	1
5	**ANN**	0.14	0.83	0.14	0.83	0.167	0.02	0.17	0.48
	NB	0.47	0.88	0.53	0.86	0.21	0.10	0.18	0.67
	DT	1	1	1	1	0.15	0.15	0.15	1

Table VI
Comparision of the Existing and the proposed Work

Classification	*Data*	*Approach*	*Algorithm (method)*	*Accuracy (%)*
DMD & BMD	Gene Sequences	MLPA – Laboratory	mPCR	75
DMD	Gene Sequences	MLPA – Laboratory	DHPLC	86
LGMD	Family Details	Machine Learning	ANN	98
6 types of MD	Micro array – Protein protein Interaction	Machine Learning	MSVM	86
FSHD	Microarray	Machine Learning	SVM	84.65
Gene type Classification	HLA Gene	Machine Learning	SVM	99.3
Virus Type Classification	HCV Virus	Machine Learning	SVM	100
5 Types of MD	Synonymous and Non – synonymous mutated gene sequences	Machine Learning	Decision Tree	100

4. DISCUSSION

The aim of this research work is to identify the proper features for classifying and building the classifier for effectiveness. As each disease has its own character the vicinity of this work depends on capturing the attributes of the gene sequence that differentiates one type of muscular dystrophy disease from another. Based on the features of missense, nonsense and silent mutations an attempt is made to build a model for predicting the type of muscular dystrophy. The missense and nonsense mutational feature comprises of annotation features, structural features and alignment features. RSCU (Relative synonymous codon usage) is calculated in the mutated gene sequences to accompany silent mutations. From the Table V it is observed that the statistics is high for decision tree than other algorithms. The positive prediction value and negative prediction value also gives a high score value for decision tree learning. The graph in Fig.3 shows the prediction value is balanced for decision tree algorithm. Prevalence is the proportion of particular disease at a specified point in time. The sensitivity measure or recall depends on prevalence and where the specificity is independent of prevalence. In decision tree based model the prevalence measures are stabilized for all classes that is exposed in Fig. 5. The detection rate and detection prevalence also depends on the prevalence measure which is also stabilized. Overall, in Fig. 6. it is made known that the balance accuracy measure is also eminent in decision tree when measured with other algorithms. The sensitivity and specificity measure for all classes is high in decision tree when compared with other learning methods. ROC curve in Fig.7. is an evidence which shows that the decision tree attains elevated sensitivity and specificity. The line that breaks at 1 shows more sensitivity and specificity. Also the experiment proves that the features designed for building the classifier are more appropriate and suitable for disease identification.

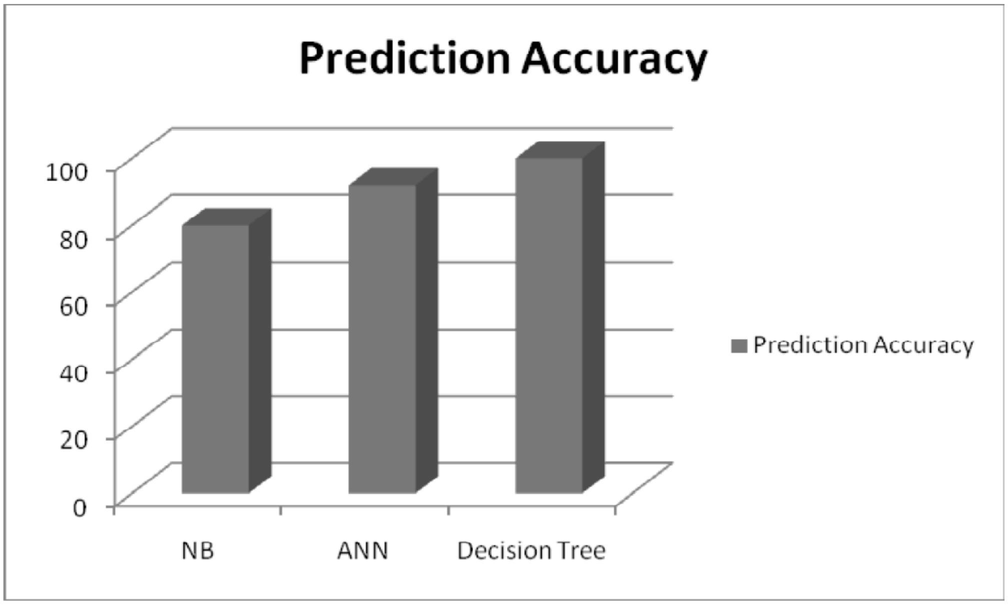

Figure 3: Comparison of Prediction accuracy on the three classification algorithms

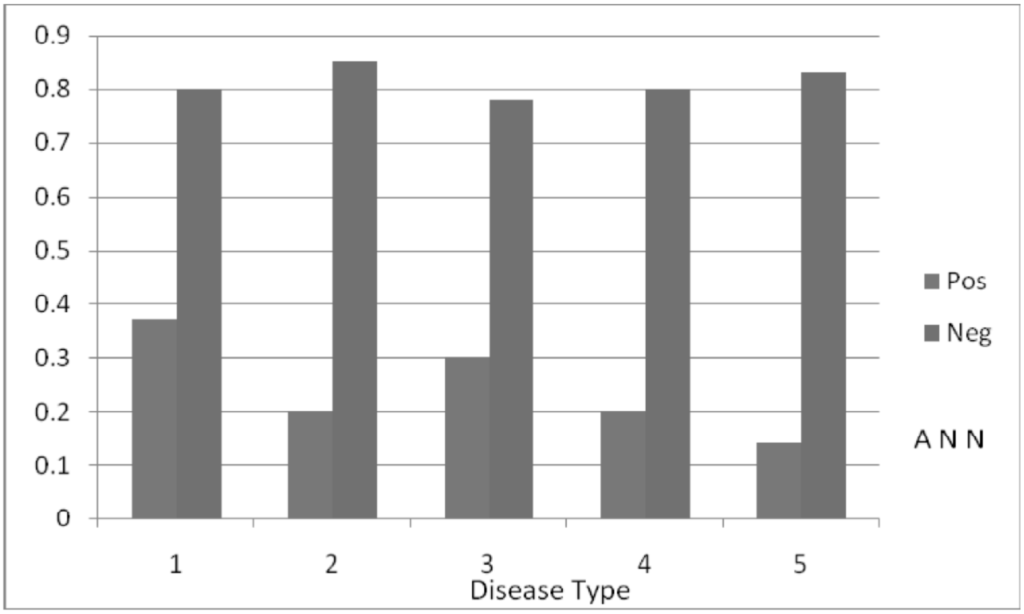

Figure 4.a

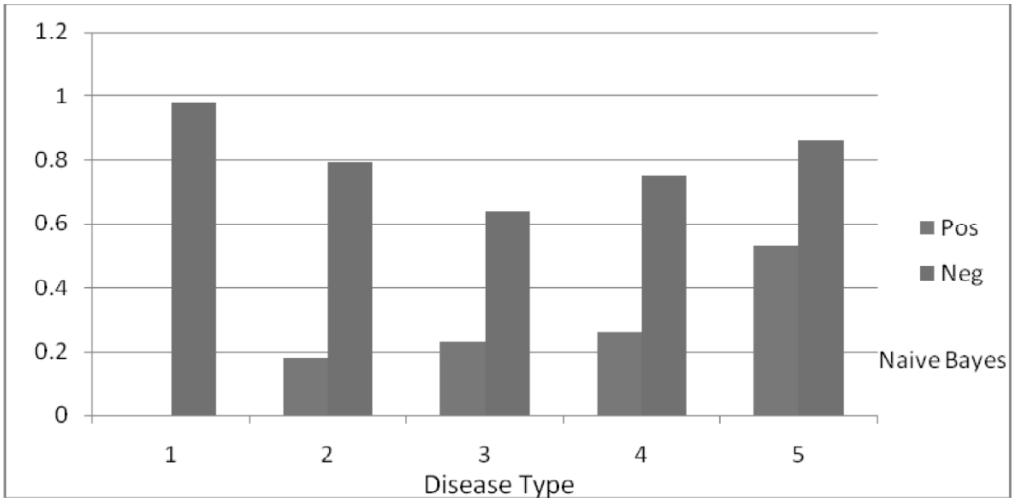

Figure 4.b

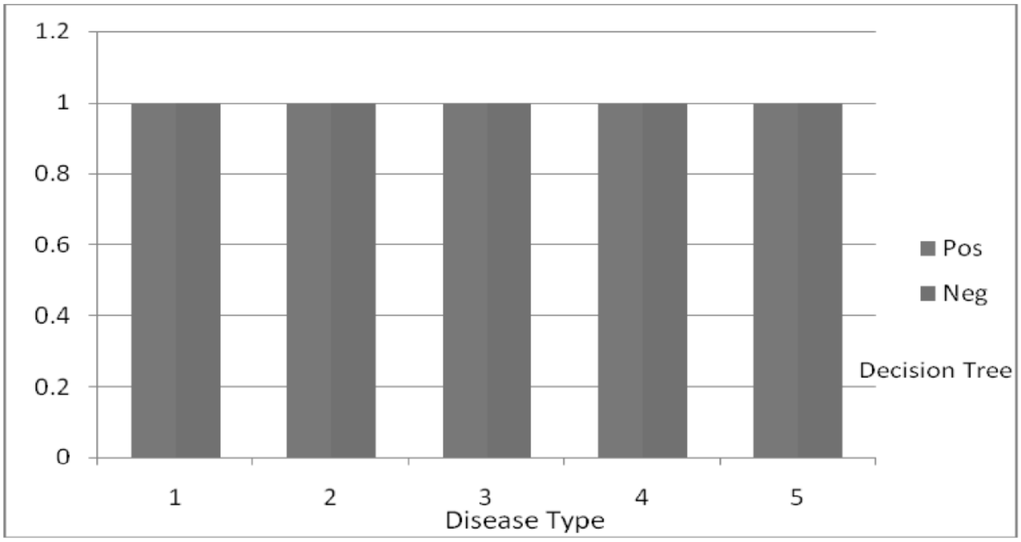

Figure 4.c

Figure 4: Positive and negative prediction value for all the classes. Fig. 4.a the values in ANN are depicted, in Fig. 4.b the values for Naïve Bayes and in Fig. 4.c the chart is shown for decision tree

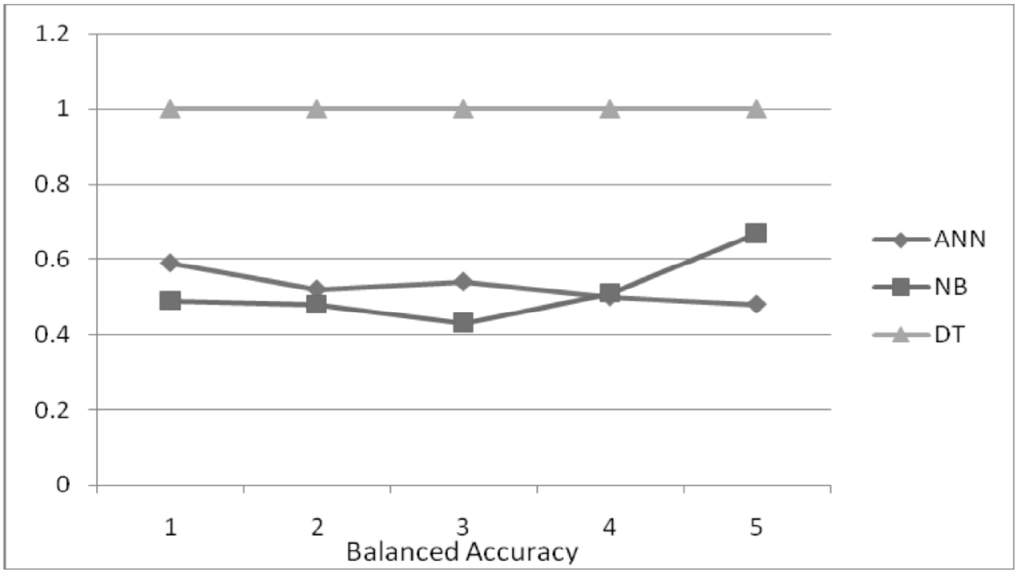

Figure 5: Prevalence measure for the classification algorithms

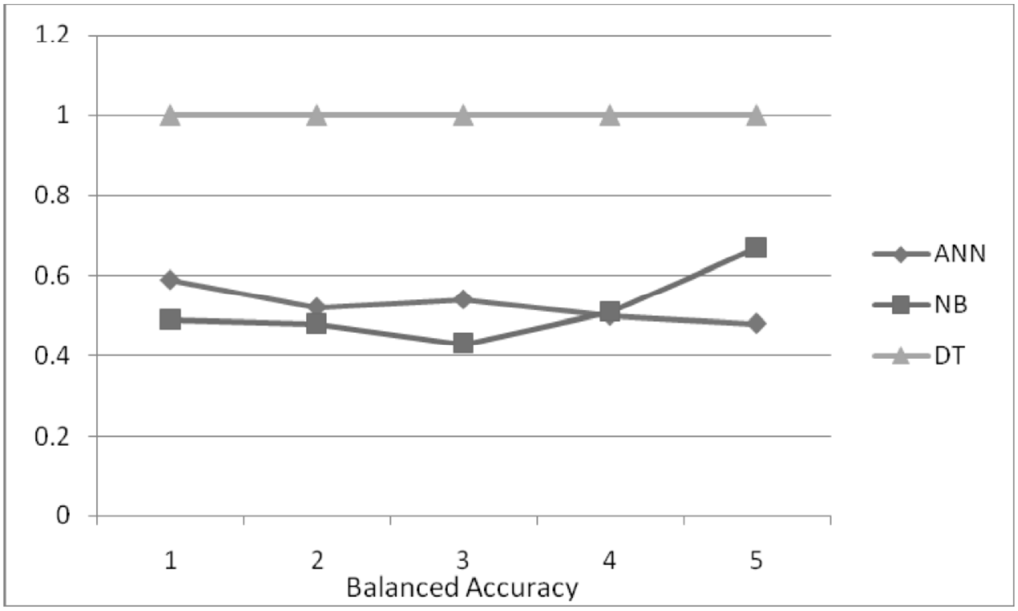

Figure 6: Overall balance accuracy of ANN, NB and DT

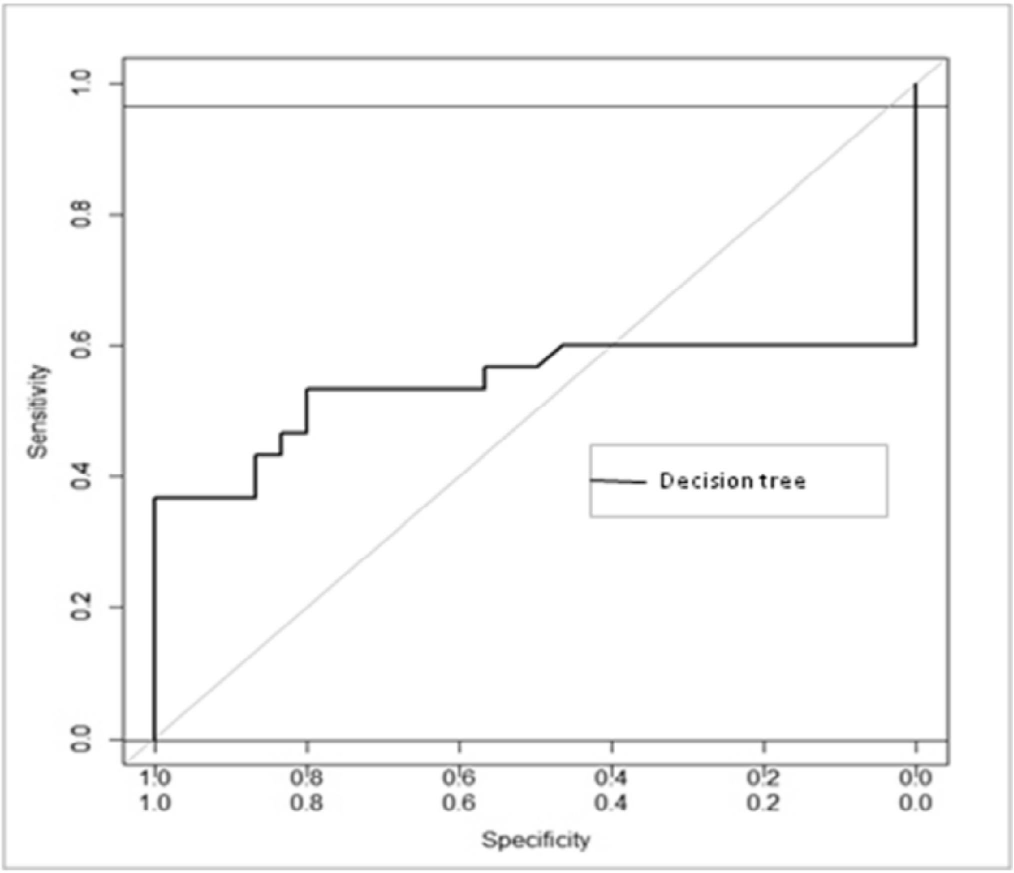

Figure 7: ROC curve analysis of sensitivity and specificity for decision tree algorithm

In this experiment the mutation spectrum accompanies all types of muscular dystrophy diseases for modeling and therefore the task of full sequencing is eliminated. This approach generalizes the disease identification task an automated practice that can be applied in identifying any kind of genetic disease. Also the prediction model is more effective and reliable since it is generated based on intelligent hints collected from mutated gene sequences.

5. CONCLUSION

Muscular dystrophy disease identification work is the problem of learning multiclass classification system that can suits in bioinformatics environment to identify the disease effectively. Currently, this problem has not been broadly studied in the literature, and existing approaches are either restricted to a small number of classes due to computational issues or insufficient data. The proposed model relies on two main ideas. The first idea is to design the discriminative features from the mutated gene sequences to build a model for identifying the type of the genetic disorder. The second idea is to capture synonymous and non– synonymous SNV's from the generated sequences and to classify the type of disease. The experiments led on the diseased gene sequences and assessed with evaluation method on the model built, show that our method is valuable than existing disease identification procedures with respect to significant features. Furthermore, when the supervised learning algorithms are applied the decision tree classifier outperforms other algorithms in prediction. As the nature of the application demands in more accurate disease prediction through mutated gene sequences, it is found that applying the extracted features in machine learning approach is significant to identify the type of muscular dystrophy disease.

NOTE

1. *http://www.hgmd.org*

REFERENCES

[1] Inusha Panigrahi, Balraj Mittal, "Carrier Detection and Prenatal Diagnosis in Duchenne/Becker Muscular Dystrophy", Indian Pediatrics 2001; 38: 631-639.

[2] Leigh B. Waddell, *et al.*, "Diagnosis of the Muscular Dystrophies"Institute for Neuroscience and Muscle Research,Children's Hospital at Westmead and Discipline of Paediatrics & Child Health,University of sydney, Australia.

[3] Lenka Fajkusova, ZdeneIk LukasIb, Miroslava Tvrdoakova a, Viera Kuhrova a,Jirioa Haajekb, Jirioa Fajkusc, Novel dystrophin mutations revealed by analysis of dystrophin mRNA:alternative splicing suppresses the phenotypic effect of a nonsense mutation Neuromuscular Disorders 11 (2001).

[4] Kevin M. Flanigan, M.D.The Muscular Dystrophies, Thieme Medical Publishers ISSN 0271-8235 Seminars in Neurology Vol. 32 No. 3/2012.

[5] Zubrzycka-Gaarn EE, Bulman DE, Karpati G, et al. The Duchenne muscular dystrophy gene product is localized in sarcolemma of human skeletal muscle. Nature 1988;333(6172): 466±469.

[6] Katharine Bushby, *et al.*, "Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management" The Lancet, November 30, 2009 DOI:10.1016/S1474-4422(09)70271-6.

[7] Anne Helbling-Leclerc, GiseÁle Bonne, and Ketty Schwartz, "Emery-Dreifuss muscular dystrophy" European Journal of Human Genetics (2002).

[8] Ma WJ1, Hashii M, *et al.*, "Non-synonymous single-nucleotide variations of the human oxytocin receptor gene and autism spectrum disorders: a case-control study in a Japanese population and functional analysis" Molecular Autism, 2013.

[9] Shuai Zeng, Jing Yang, *et al.*, "EFIN: predicting the functional impact of nonsynonymous single nucleotide polymorphisms in human genome", BMC Genomics, 2014.

[10] *http://www.ncbi.nlm.nih.gov/books/NBK21578*

[11] Kann, M.G.: Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefings in Bioinformatics 11, 96–110 (2009).

[12] Tranchevent, L.-C., *et al.*: A guide to web tools to prioritize candidate genes. Briefings in Bioinformatics 12, 22–32 (2010).

[13] Aleksandra Nadaj-Pakleza, *et al.*, "The role of skeletal muscle biopsy in the diagnosis of neuromuscular disorders" 2010 Polish Society of Neurology and the Polish Association of Neurosurgeons, Elsevier.

[14] *http://evolution.berkeley.edu/evolibrary/article/ mutations_01*

[15] KN North and KJ Jones., "Diagnosing childhood muscular dystrophies. Journal of Paediatrics and Child Health".

[16] Roberts *et al*. “Point mutations in the dystrophin gene “, Vol.9 March 1992 Genetics.

[17] Chen Chen, Hongwei Ma, “Screening of Duchenne Muscular Dystrophy (DMD) Mutations and Investigating Its Mutational Mechanism in Chinese Patients”, PLOS One 2014.

[18] Bennett RR1, Schneider HE *et al.*, “Automated DNA mutation detection using universal conditions direct sequencing: application to ten muscular dystrophy genes”, BMC Genetics 2009.

[19] Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene innormal and affected individuals. Cell 1987; 50:509±517.

[20] Dr. Mohini Joshi, Dr.Deshpande J.D, “ POLYMERASE CHAIN REACTION: METHODS, PRINCIPLES AND APPLICATION”, International Journal of Biomedical Research, 2011.

[21] Hyeyoung Lee, Dong Wook Jekarl, Joonhong Park, Hyojin Chae, Myungshin Kim,Yonggoo Kim, and Jong in Lee Identification of DMD Mutation in Korean Siblings Using Full Gene Sequencing.

[22] Felix F.Gonzalez-Navarro et.al,”Effective Classification and Gene Expression Profiling for the Facioscapulohumeral Muscular Dystrophy” , PLoS ONE 2013.

[23] Madhuri R. Hegde1, Ephrem L.H. Chin1, Jennifer G. Mulle1, David T. Okou1, Stephen T. Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia Microarray-based mutation detection in the dystrophin gene.

[24] Satoru Noguchi1, Toshifumi Tsukahara, Masako Fujita, Rumi Kurokawa, cDNA microarray analysis of individual Duchenne muscular dystrophy patients Human Molecular Genetics, 2003, Vol. 12, No. 6.

[25] Catherine T. Falk, *et al.* “Using Neural Networks as an Aid in the Determination of Disease Status:Comparison of Clinical Diagnosis to Neural-Network Predictions in a Pedigree with Autosomal Dominant Limb-Girdle Muscular Dystrophy”,Am. J. Hum. Genet. 62:941–949, 1998.

[26] Jianmin Ma, Minh N. Nguyen, Gavyn W. L.Pang, and Jagath C. Rajapakse, “Gene Classification using Codon Usage and SVMs”, IEEE, 2005.

[27] C. M. Nisha, Bhasker Pant, and K. R. Pardasani,”SVM model for classification of genotypes of HCV using Relative Synonymous Codon Usage” Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604.Online ISSN 2278 – 6007 Vol 3, Issue 3, 2012, pp 357-363.

[28] Peter D. Stenson • Matthew Mort • Edward V. Ball • Katy Shaw • Andrew D. Phillips • David N. Cooper The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine July 2013.